AFRL-IF-RS-TR-2005-195
Final Technical Report
May 2005

ENCODING COOPERATIVE DNA CODES

Anthony J. Macula, Inc.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2005-195 has been reviewed and is approved for publication

APPROVED: /s/
THOMAS RENZ
Project Engineer

FOR THE DIRECTOR: /s/
JAMES A. COLLINS, Acting Chief
Advanced Computing Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 2005	**Final** **Jan 03 – Jan 05**

4. TITLE AND SUBTITLE

ENCODING COOPERATIVE DNA CODES

5. FUNDING NUMBERS

C - F30602-03-C-0059
PE - 61102F
PR - EIDN
TA - AC
WU - 01

6. AUTHOR(S)

Anthony J. Macula

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Anthony J. Macula, Inc.
36 Westview Cr
Geneseo NY 14454

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/IFTC
26 Electronic Parkway
Rome NY 13441-4514

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

AFRL-IF-RS-TR-2005-195

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: Thomas Renz/IFTC/(315) 330-3423 Thomas.Renz@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The primary goal of this research was the development of an enabling technology for DNA computing. It is focused on the construction of a biomolecular architecture designed to employ new algorithmic paradigms based on the massively parallel computational power of DNA hybridization. The intent is to develop a computing basis to eventually overcome the exponential time complexity of many discrete math problems so that they can be solved in linear real time. Many of these computationally hard (NP) problems are critical to logistics, scheduling and security. In this way, this research addresses computational, national security and knowledge acquisition challenges of the Air Force. DNA codewords are structural and information building blocks in biomolecular computing and other biotechnical applications that employ DNA hybridization assays. Thermodynamic distance functions are important components in the construction of DNA codes. We introduce new matrics for DNA code design that captures key aspects of the nearest neighbor thermodynamic model for hybridized DNA duplexes. One version of our metric gives the maximum number of stacked pairs of hydrogen bonded nucleotide base pairs that can be present in any secondary structure in a hybridized DNA duplex without pseudoknots. We introduce the concept of (t-gap) block isomorphic subsequences to describe new string metrics that are similar to the weighted Levenshtein insertion-deletion metric. We show how our new distances can be calculated by a generalization of the folklore longest common subsequence dynamic programming algorithm. We give a Varshamov-Gilbert like lower bound on the size of some of codes using our distance functions as constraints. We also discuss software implementation of our DNA code design methods.

14. SUBJECT TERMS

DNA Computing, Synthetic DNA, Biomolecular Computing

15. NUMBER OF PAGES 33

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Figures

1. Summary

The primary goal of this research was the development of an enabling technology for DNA computing. It is focused on the construction of a biomolecular architecture designed to employ new algorithmic paradigms based on the massively parallel computational power of DNA hybridization. The intent is to develop a computing basis to eventually overcome the exponential time complexity of many discrete math problems so that they can be solved in linear real time. Many of these computationally hard (NP) problems are critical to logistics, scheduling and security. In this way, this research address computational, national security and knowledge acquisition challenges of the Air Force. In this report we:

1. Give new metrics for cooperative DNA code design that capture key aspects of the nearest neighbor thermodynamic model for hybridized DNA duplexes.
2. Show how DNA computing can be applied to the identification of independent sets in a graph.
3. Show how our software uses our new metrics to construct a biomolecular computing architecture to address the identification of independent sets in a graph.

2. Introduction

DNA codewords are structural and information building blocks in biomolecular computing and other biotechnical applications that employ DNA hybridization assays. Thermodynamic distance functions are important components in the construction of DNA codes. We introduce new metrics for DNA code design that capture key aspects of the nearest neighbor thermodynamic model for hybridized DNA duplexes. One version of our metric gives the maximum number of stacked pairs of hydrogen bonded nucleotide base pairs that can be present in any secondary structure in a hybridized DNA duplex

without pseudoknots. We introduce the concept of (t-gap) block isomorphic subsequences to describe new string metrics that are similar to the weighted Levenshtein insertion-deletion metric. We show how our new distances can be calculated by a generalization of the folklore longest common subsequence dynamic programming algorithm. We give a Varshamov-Gilbert like lower bound on the size of some of codes using our distance functions as constraints. We also discuss software implementation of our DNA code design methods.

In this paper, all variables are nonnegative integers unless otherwise stated. $[n]$ denotes the set $\{0,\ldots,n-1\}$ and (n) denotes the sequence $1,2,\ldots,n$. Given two sequences α and β, we write $\alpha \prec \beta$ if and only if α is a subsequence of β. The length of sequence α is denoted by $|\alpha|$. We call $\alpha \prec (n)$ a *string* if and only if it is a subsequence of consecutive integers, e.g., $\alpha = i, i+1, \ldots, i+k$. For $a \le b$, we use the notation $[a,b]$ for the string of integers between and including a and b. If $a = b$, we sometimes write $[a]$ for $[a,b]$. When we write $\sigma = [a_1,b_1],[a_2,b_2],\ldots,[a_i,b_i],\ldots[a_k,b_k]$ where $a_i \le b_i < a_{i+1}$, we mean $\sigma = a_1, a_1+1,\ldots,b_1,a_2,a_2+1,\ldots b_2,\ldots,a_i,a_i+1,\ldots b_i,\ldots a_k,a_k+1,\ldots b_k$. For $\sigma \prec (n)$, $\tau \prec (m)$ with $|\sigma| \le |\tau|$, we write $f:\sigma \to \tau$ to indicate an *increasing function* $f:\{i : i \in \sigma\} \to \{i : i \in \tau\}$. Given $\sigma = i_1,i_2,\ldots,i_k$ and $f:\sigma \to \tau$, we define $f(\sigma) \equiv f(i_1),f(i_2),\ldots,f(i_k)$. If $|\sigma| = |\tau|$, then $f:\sigma \to \tau$ is unique. We let $[4]^n$ denote the set of sequences of length n with entries in $[4]$. We identify {0,1,2,3} with the symbols for DNA bases {A,C,G,T}. Thus sequences in $[4]^n$ are all DNA sequences of length n. For $x = x_1,\ldots,x_n$ with $x \in [4]^n$ and $\sigma = i_1,i_2,\ldots,i_k$ where $\sigma \prec (n)$, we let $x_\sigma \prec x$ be the subsequence $x_{i_1},x_{i_2}\ldots,x_{i_k}$. Given a non-negative real-valued function, Ω, on $[q]$, we define $\|x_\sigma\|_\Omega \equiv \sum_{i\in\sigma} \Omega(x_i)$.

3. Methods, Assumption, Procedures

3.1 Block Isomorphic Subsequences

Definition 1. *For* $\sigma \prec (n)$, *a substring* $\beta \prec \sigma$ *is called a block of* σ *if* β *is not subsequence of any substring* α *of* σ *with* $\beta \neq \alpha$. *A subsequence* $x_\beta \prec x_\sigma$ *is called a block of* x_σ *if* β *is a block of* σ. *For* $\sigma \prec (n)$, *let* $(\beta_i(\sigma))$ *be the sequence of blocks of* σ, *where each element of* $\beta_i(\sigma)$ *is less than every element of* $\beta_{i+1}(\sigma)$. *Let* $b_i(\sigma)$ *and* $B_i(\sigma)$ *be the left and right endpoints of* $\beta_i(\sigma)$. *When the context is clear, we just write* b_i, B_i *and* β_i *respectively. When we write* $\sigma = ([b_i, B_i]) = (\beta_i)$, *we say that it is the block representation of* σ. *When we write* $x_\sigma = x_{(\beta_i)}$ *we say that it is the block representation of* x_σ.

Block representations are unique. Let $x \in [4]^{14}$ and let $\sigma \prec (14)$. Suppose $x = 2,0,1,2,2,3,0,3,2,0,0,1,3,2$ and $\sigma = 2,3,4,7,9,10,13,14$. Then $x_\sigma = 0,1,2,0,2,0,3,2$ and the block representations are: $\sigma = [2,4],\ [7,7],\ [9,10],\ [13,14]$ and $x_\sigma = x_{[2,4],[7,7],[9,10],[13,14]}$.

Definition 2. *For* $2 \leq t \leq n-1$, *we define*

$$G_t(n) \equiv \{\sigma \prec (n) : b_{i+1}(\sigma) - B_i(\sigma) \geq t\}.$$

We call $G_t(n)$ *the set of t-gap sequences of (n).*

Note that $\sigma \prec (n) \Rightarrow \sigma \in G_2(n)$ and $2 \leq t_1 \leq t_2 \leq n-1 \Rightarrow G_{t_2}(n) \subseteq G_{t_1}(n)$.

Definition 3. *Let* $\sigma \prec (n), \quad \tau \prec (m)$ *with* $|\sigma| = |\tau|$. *Let* $f : \sigma \to \tau$ *be unique. We say that* σ *and* τ *are block isomorphic and write* $\sigma \cong \tau$ *if:* $\alpha \prec \sigma$ *is a string* $\Leftrightarrow f(\alpha) \prec \tau$ *is a string. For* $x \in [q]^n$, $y \in [q]^m$ *we say that* x_σ *and* y_τ *are block isomorphic, denoted by* $x_\sigma \cong y_\tau$, *if* $x_\sigma = y_\tau$ *and* $\sigma \cong \tau$.

Proposition 1. *Suppose* $\sigma \cong \tau$ *and* $f : \sigma \to \tau$ *is unique. Then* $\beta \prec \sigma$ *is a block in* σ *if and only if* $f(\beta)$ *is a block in* τ .

Definition 4. *Let* $2 \le t \le \min(n,m) - 1$ *and suppose* $\sigma \prec (n), \quad \tau \prec (m)$ *with* $|\sigma| = |\tau|$. *We say that* σ *and* τ *are t-gap block isomorphic and write* $\sigma \overset{}{\underset{t}{\cong}} \tau$ *if and only if* $\sigma \in G_t(n), \quad \tau \in G_t(m)$ *and* $\sigma \cong \tau$. *For* $x \in [q]^n$, $y \in [q]^m$,*we say that* x_σ *and* y_τ *are t-gap block isomorphic, denoted by* $x_\sigma \overset{}{\underset{t}{\cong}} y_\tau,$ *if and only if* $x_\sigma = y_\tau$ *and* $\sigma \overset{}{\underset{t}{\cong}} \tau$. *We say that* x *and* y *have a common t-gap block isomorphic subsequence if and only if there are* $\sigma \prec (n), \quad \tau \prec (m)$ *with* $x_\sigma \overset{}{\underset{t}{\cong}} y_\tau$. *Note that* $\sigma \cong \tau \Leftrightarrow \sigma \overset{}{\underset{2}{\cong}} \tau$ *and* $\sigma \overset{}{\underset{t_2}{\cong}} \tau \Rightarrow \sigma \overset{}{\underset{t_1}{\cong}} \tau$ *when* $t_1 \le t_2$. *So, we just write* $\sigma \cong \tau$ *and* $x_\sigma \cong y_\tau$ *to denote* $\sigma \overset{}{\underset{2}{\cong}} \tau$ *and* $x_\sigma \overset{}{\underset{2}{\cong}} y_\tau$ *respectively* .

Example 1. *Let* $x,y,z,w \in [4]^{13}$ *and* $\sigma_i \prec (13)$ *with x=1, 1, 2, 0, 2, 3, 3, 0, 1, 1, 2, 0, 1; y=2, 0, 2, 3, 3, 0, 1, 1, 1, 1, 2, 0, 3; z=3, 2, 0, 2, 1, 1, 1, 1, 1, 0, 3, 2, 0; w=1, 1, 1, 2, 0, 3, 2, 0, 2, 0, 3, 3, 3* . *Let* σ_1 = *[3,5], [8,8], [11,12]* ; σ_2 = *[1,3], [6,6], [11,12]* ; σ_3 = *[2,4], [10,10], [12,13]* ; σ_4 = *[4,5], [7,10]* . *Then* $x_{\sigma_1} = y_{\sigma_2} = z_{\sigma_3} = w_{\sigma_4} = 2,0,2,0,2,0.$ *Since* $\sigma_1 \cong \sigma_2 \cong \sigma_3 \ncong \sigma_4,$ *we have that* $x_{\sigma_1} \cong y_{\sigma_2} \cong z_{\sigma_3} \ncong w_{\sigma_4}.$ *Since* $\sigma_1, \sigma_2 \in G_3(n)$ *and* $\sigma_3 \notin G_3(n),$ *we have that*

$x_{\sigma_1} \underset{3}{\cong} y_{\sigma_2} \underset{3}{\ncong} z_{\sigma_3}$.

3.2 Block Insertion-Deletion Codes

Definition 5. *For* $x,y \in [q]^n$ *, we define:* $\rho_{\Omega,q}(x,y) \equiv \max\{\|z\|_\Omega : z \prec x \text{ and } z \prec y\}$ *and* $L_{\Omega,q}(x,y) \equiv \min(\|x\|_\Omega, \|y\|_\Omega) - \rho_\Omega(x,y)$. *We say that* $\rho_\Omega(x,y)$ *is the maximum weight of a common subsequence to* x *and* y *and* $L_\Omega(x,y)$ *is called the weighted Levenshtein insertion-deletion distance.* $L_\Omega(x,y)$ *is a metric. When* $\|x_\sigma\|_\Omega = |x_\sigma|$, *we write* $L(x,y)$ *for* $L_\Omega(x,y)$.

Definition 6. *For* $2 \le t \le n-1$ *and* $x,y \in [q]^n$ *. We define:*

$$\phi^t_{\Omega,q}(x,y) \equiv \max\{\|x_\sigma\|_\Omega : x_\sigma \underset{t}{\cong} y_\tau\}.$$

$$\Phi^t_{\Omega,q}(x,y) \equiv \min(\|x\|_\Omega, \|y\|_\Omega) - \phi^t_{\Omega,q}(x,y).$$

When the context for q *is clear, we simply write* $\Phi^t_\Omega(x,y)$ *and* $\phi^t_\Omega(x,y)$. *We say that* $\phi^t_\Omega(x,y)$ *is the weight of the longest common t-gap block subsequence of* x *and* y . *When* $\|x_\sigma\|_\Omega = |x_\sigma|$, *we write* $\Phi^t(x,y)$ *and* $\phi^t(x,y)$ *for* $\Phi^t_\Omega(x,y)$ *and* $\phi^t_\Omega(x,y)$ *respectively. For* $t = 1$, *we define* $\phi^1_\Omega(x,y) \equiv L_\Omega(x,y)$ *and* $\phi^1(x,y) \equiv L(x,y)$.

Proposition 2. $\Phi^t_\Omega(x,y)$ *is a metric on* $[q]^n$.

Definition 7. *A t-gap block insertion-deletion q-ary code of weighted* Ω *distance d is a subset* C, *of* $[q]^n$, *such that:* $x \ne y \in C \Rightarrow \Phi^t_\Omega(x,y) \ge d$.

Theorem 1. *Let* $\|x_\sigma\|_\Omega = |x_\sigma|$. *In* $[q]^n$ *, there is a 2-gap block insertion-deletion* C *of* $d = n - k$ *with*

$$|C| \geq q^k \left(\sum_{j=1}^{k} \binom{k-1}{j-1} \binom{n-k+1}{j}^2 \right)^{-1}.$$

4. Results, Discussion

4.1 Computing $\phi_\Omega^t(x,y)$

For $1 \leq m < n$, consider the string $[m+1,n]$. For $x,y \in [q]^n$, let suf(x,y) be the length of the longest common suffix between x and y. Then $suf(x,y) = 0$ if $x_n \neq y_n$ and $suf(x,y) = n-m$ if $x_{[m+1,n]} = y_{[m+1,n]}$ and $x_m \neq y_m$. For $x \in [q]^n$, we have that $x_{[1,i]}$ is the first i entries of x and $x_{[1,n]} = x$.

Proposition 3. *Let* $1 \leq t \leq n-1$. *For* $x,y \in [q]^n$ *and* $t < i,j \leq n$, *define* $M_{\Omega,i,j}^t \equiv \phi_\Omega^t(x_{[1,i]}, y_{[1,j]})$. *Let* $\omega(r) \equiv \|x_{[n-r+1,n]}\|_\Omega$ *and* $suf(x,y) = k$. *Define* $D_{\Omega,i,j}^t \equiv \max\{\omega(r) + M_{\Omega,i-r-t+1,j-r-t+1}^t : 1 \leq r \leq k\}$ *if* $k \geq 1$ *and* $D_{\Omega,i,j}^t \equiv 0$ *if* $k = 0$. *Then*

$$M_{\Omega,i,j}^t = \phi_\Omega^t(x_{[1,i]}, y_{[1,j]}) = \max\{M_{\Omega,i-1,j}^t, M_{\Omega,i,j-1}^t, D_{\Omega,i,j}^t\}.$$

When either i *or* j *is less than or equal to* t, *the initial conditions needed for the computation of* $\phi_\Omega^t(x,y)$ *are* $\phi_\Omega^t(x_{[1,i]}, y_{[1,j]}) = \|x_{[1,i]}\|_\Omega$ *if and only if* $x_{[1,i]}$ *is a substring of* $y_{[1,j]}$.

Example 2. *Let* $x = 0,1,2,3,1,3,0,1$ *and* $y = 3,0,1,3,2,0,3,1$. *Figure 1 and Figure 2 are* M^2 *and* M^3 *where* $\|x_\sigma\| = |\sigma|$. *Figure 3 and Figure 4 are* M_Ω^2 *and* M_Ω^3 *where* $\|x_\sigma\|_\Omega \equiv \sum_{i \in \sigma}(x_i + 1)$. *Below each figure, we give an example of t-gap isomorphic subsequences with* $\|x_\sigma\| = \|y_\tau\| = \varphi^t(x,y)$ $(\varphi^t(x,y).)$

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 \\ 1 & 1 & 2 & 2 & 2 & 2 & 3 & 4 \\ 1 & 1 & 2 & 2 & 2 & 2 & 3 & 4 \\ 1 & 2 & 2 & 2 & 2 & 3 & 3 & 4 \\ 1 & 2 & 3 & 3 & 3 & 3 & 3 & 4 \end{bmatrix}$$

Fig. 1: $x_{[1,2],[4,5]} \cong_3 y_{[2,3],[7,8]}$

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 & 2 & 2 & 3 \\ 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 \\ 1 & 2 & 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix}$$

Fig. 2: $x_{[1,2],[8,8]} \cong_3 y_{[2,3],[8,8]}$

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 3 & 3 & 3 & 3 & 3 \\ 0 & 1 & 3 & 3 & 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 5 & 5 & 5 & 7 & 7 \\ 4 & 4 & 4 & 5 & 5 & 5 & 7 & 9 \\ 4 & 4 & 4 & 8 & 8 & 8 & 9 & 9 \\ 4 & 5 & 5 & 8 & 8 & 8 & 9 & 9 \\ 4 & 5 & 7 & 8 & 8 & 9 & 9 & 10 \end{bmatrix}$$

Fig. 3: $x_{[4],[6],[8]} \cong y_{[1],[4],[7]}$

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 3 & 3 & 3 & 3 & 3 \\ 0 & 1 & 3 & 3 & 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & 4 & 4 & 4 & 5 & 7 \\ 4 & 4 & 4 & 6 & 6 & 6 & 7 & 7 \\ 4 & 5 & 5 & 6 & 6 & 6 & 7 & 7 \\ 4 & 5 & 7 & 7 & 7 & 7 & 7 & 7 \end{bmatrix}$$

Fig. 4: $x_{[1],[4,5]} \cong_3 y_{[2],[7,8]}$

4.2 Sequences of t-Strings

In this section, we apply the results of previous sections to sequences of strings of length t (with particular attention to $t = 2$) that naturally arise from $x \in [q^n]$. The goal is to then apply these results to the modeling of DNA hybridization distances.

Definition 8. *For* $\sigma,\tau \prec (n)$ *and* $1 \le t \le n-1$, *let* $\sigma^t \prec \sigma$ *be defined by:* $i \in \sigma^t \Leftrightarrow [i, i+t-1] \prec \sigma$. *If* $|\sigma| = |\tau|$, *then for the unique* $f : \sigma \to \tau$, *we define* σ^t_τ *as:* $i \in \sigma^t_\tau \Leftrightarrow [i, i+t-1] \prec \sigma$ *and* $[f(i), f(i)+t-1] \prec \tau$. *We define* $\tau^t_\sigma \prec \tau$ *by* $\tau^t_\sigma \equiv f(\sigma^t_\tau)$.

Example 3. *Let* $\sigma,\tau \prec (16)$ *be given in their block representations with* $\sigma = [1,4]$, $[7,10]$, $[12,15]$ *and* $\tau = [2,8]$, $[12,16]$. *Then* σ^2 , τ^2 , σ^2_τ *and* τ^2_σ *are:* $\sigma^2 = [1,3]$, $[7,9]$, $[12,14]$; $\tau^2 = [2,7],[12,15]$; $\sigma^2_\tau = [1,3],[7,8],[12,14]$; $\tau^2_\sigma = [2,4],[6,7],[13,15]$.

Then σ^3 , τ^3 , σ^3_τ *and* τ^3_σ *are:* $\sigma^3 = [1,2],[7,8],[12,13]$; $\tau^3 = [2,6]$, $[12,14]$; $\sigma^3_\tau = [1,2],[7,7],[12,13]$; $\tau^3_\sigma = [2,3],[6,6],[13,14]$. *A careful inspection of* σ^2_τ , τ^2_σ *and* σ^3_τ, τ^3_σ *demonstrates the general result that* $\sigma^t_\tau \underset{t}{\cong} \tau^t_\sigma$.

For $x,y \in [q]^n$ with $x_\sigma = y_\tau$, we have that $i \in \sigma^t_\tau$ if and only if i is the first index in a common t-string , $x_{[i,i+t-1]} = y_{[f(i),f(i)+t-1]}$, of the common subsequence $x_\sigma = y_\tau$ of x and y . Thus $|\sigma^t_\tau|$ *is the number of common t-strings that occur in the common subsequence* $x_\sigma = y_\tau$ *of* x and y . In particular, $|\sigma^2_\tau|$ *is the number of common 2-strings that occur in the common subsequence* $x_\sigma = y_\tau$ *of* x and y . This measure is of interest to us because when two DNA strands have a secondary structure in a duplex, the thermodynamic weight (e.g., free energy) of *nearest neighbor stacked pairs* of that secondary structure is a measure (*not the measure*) of the thermodynamic stability of the duplex with the given secondary structure. Since every secondary structure in the DNA duplex $x : \overleftarrow{\overline{y}}$ between x and the complement, $\overline{y}$, of y corresponds to a common subsequence, $x_\sigma = y_\tau$, between x and y, we have that $|\sigma^2_\tau|$ gives us the number of nearest neighbor stacked pairs in the $x : \overleftarrow{\overline{y}}$ duplex with the secondary

structure associated with $x_\sigma = y_\tau$. In general, $|\sigma_\tau^t|$ gives us the number of *common t-stems* in the $x : \overleftarrow{y}$ duplex with the secondary structure associated with $x_\sigma = y_\tau$. We now show how we compute the "total weight" of the common 2-strings that occur in the common subsequence $x_\sigma = y_\tau$ of x and y .

Definition 9. *Suppose* $2 \le t \le n-1$. *Given a string* $[a,b] \prec (n)$ *and* $x \in [q]^n$ *, let* $d_q(x_{[a,b]})$ *be the unique number in* $[q^t]$ *whose q-ary decimal representation is* $x_a x_{a+1} x_{a+2} \ldots x_b$. *For* $x \in [q]^n$ *, let* $x^{(t)} \in [q^t]^{n-t}$ *be defined as* $x^{(t)} \equiv (d_q(x_{[i,i+t-1]}))_{i=1}^{n-t+1}$. *For example, if* $x = 2, 3, 3, 0, 3, 0, 2, 2, 1, 1, 2, 0, 2 \in [4]^{13}$ *, then* $x^{(2)} = 11, 15, 12, 3, 12, 2, 10, 9, 5, 6, 8, 2 \in [16]^{12}$ *and* $x^{(3)} = 47, 60, 51, 12, 50, 10, 41, 37, 22, 24, 34 \in [64]^{11}$.

Definition 10. *Suppose* $2 \le t \le n-1$. *Let* Ω *be a weight function on* $[q^t]$. *Then for* $x,y \in [q]^n$, *we define:* $\psi_\Omega^t(x,y) \equiv \max\{\left\| x_{\sigma_\tau^t}^{(t)} \right\|_\Omega : x_\sigma = y_\tau\}$ *and* $\Psi_\Omega^t(x,y) \equiv \min\left(\|x^{(t)}\|_\Omega, \|y^{(t)}\|_\Omega\right) - \psi_\Omega^t(x,y)$. *If* $\|x_\sigma\|_\Omega = |\sigma|$, *then we write* $\psi^t(x,y)$ *and* $\Psi^t(x,y)$ *for* $\psi_\Omega^t(x,y)$ *and* $\Psi_\Omega^t(x,y)$ *respectively.*

Proposition 4. *Suppose* $2 \le t \le n-1$. *Let* Ω *be a weight function on* $[q^t]$. *Then for* $x,y \in [q]^n$, *we have:*

$$\psi_\Omega^t(x,y) = \phi_{\Omega,q^t}^t(x^{(t)},y^{(t)})$$

Definition 11. *A q-ary code of* Ω *weighted t-stem distance d is a subset* C *of* $[q]^n$ *such that:* $x \ne y \in C \Rightarrow \Psi_\Omega^t(x,y) \ge d$. *If* $t = 2$, *we call such a* C *a q-ary code of* Ω *weighted stacked pair distance d. Thus if* C *is a* Ω *weighted code of t-stem distance d, then:* $x \ne y \in C \Rightarrow \|x^{(t)}\|_\Omega - \phi_{\Omega,q^t}^t(x^{(t)},y^{(t)}) \ge d$ *and*

$\|y^{(t)}\|_\Omega - \phi^t_{\Omega,q^t}(x^{(t)},y^{(t)}) \geq d.$ *If* $\|x_\sigma\|_\Omega = |x_\sigma|,$ *we simply call such a* C *a t-stem code of distance d* .

From a DNA duplex point of view, with $\Omega \equiv F$ being the thermodynamic weight of the (virtual) stacked pairs of nucleotides (see Table 1,) $\|x^{(2)}\|_\Omega$ is the absolute value of nearest neighbor free energy of the duplex $x : \overleftarrow{x}$. Thus, if C is a (A,C,G,T) quaternary code of F weighted stacked pair distance d, then $x \neq y \in C$ implies that the thermodynamic stability of each of the duplexes $x : \overleftarrow{x}$ and $y : \overleftarrow{y}$ is at least "d greater than" the thermodynamic stability of the duplex $x : \overleftarrow{y}.$ Moreover, if C is a (A,C,G,T) quaternary t-stem code of distance d, then $x \neq y \in C$ implies $x : \overleftarrow{x}$ and $y : \overleftarrow{y}$ each have at least "d more" common t-stems than are in any secondary structure for duplex $x : \overleftarrow{y}.$ The main point of application is that $\psi^2_F(x,\overline{y})$ is a measure of the nearest neighbor stability of the DNA duplex $x : \overleftarrow{y}$ and $\psi^t(x,y)$ is the maximum number of t-stems that can form in any secondary structure of $x : \overleftarrow{y}.$ For a 2-stem code C of distance $d = (n-1) - k$, we have for $x \neq y \in C,$ that the *maximum number* of stacked pairs in a secondary structure of the duplex $x : \overleftarrow{y}$ is at most k while the number of stacked pairs in each of the $x : \overleftarrow{x}$ and $y : \overleftarrow{y}$ duplex is $n-1$.

Theorem 2. In $[q]^n,$ there is a 2-stem code, $C,$ of distance $d = (n-1) - k$ with:

$$|C| \geq q^k \left(\sum_{j=1}^{k} q^{-j} \binom{k-1}{j-1} \binom{n-k}{j}^2 \right)^{-1}.$$

5. Conclusions

5.1 Applications to DNA Hybridization Distance Modeling

Single strands of DNA are, abstractly, (A,C,G,T) -quaternary sequences, with the four letters denoting the respective nucleic acids. DNA sequences are oriented; e.g., $5'AACG3'$ is distinct from $5'GCAA3'$, but it is identical to $3'GCAA5'$. The orientation of a DNA strand is usually indicated by the $5' \to 3', 3' \to 5'$ notation that reflects the asymmetric covalent linking between consecutive bases in the DNA strand backbone. In this paper, when we write DNA molecules without indicating the direction, it is assumed that the direction is $5' \to 3'$. Furthermore, DNA is naturally double stranded. That is, each sequence normally occurs with its reverse complement, with reversal denoting that two strands are oppositely directed, and with complementarity denoting that the allowed pairings of letters, opposing one another on the two strands, are $\{A,T\}$ or $\{G,C\}$ ---the canonical Watson-Crick *base pairings*. Therefore, to obtain the reverse complement of a strand of DNA, first reverse the order of the letters and then substitute each letter with its complement. For example, the reverse complement of $AACGTG$ is $CACGTT$. If $x = AACGTG$, then we let $\overline{x}$ denote it reverse complement $CACGTT$. We let $\overleftarrow{x}$ denote x listed in reverse $3' \to 5'$ order. *As DNA sequences,* x *and* $\overleftarrow{x}$ *are identical, i.e.,* $x = 5'x_1,\ldots,x_n3' \equiv \overleftarrow{x} = 3'x_n,\ldots,x_15'$

A *Waston-Crick (WC) duplex* results from joining reverse complement sequences in opposite orientations, e.g.,

$$x : \overleftarrow{\overline{x}} = \begin{matrix} 5'AACGTG3' \\ 3'TTGCAC5' \end{matrix} .$$

Whenever two, not necessarily complementary, oppositely directed DNA strands "mirror" one another sufficiently, they are capable of coalescing into a DNA duplex. The

process of forming DNA duplexes from single strands is referred to as *DNA hybridization*. The greatest energy of duplex formation is obtained when the two sequences are reverse complements of one another and the DNA duplex formed is a WC duplex. However, there are many instances when the formation non-WC duplexes are energetically favorable. In this paper, a non-WC duplex is referred to as a *cross-hybridized (CH) duplex.*

An $n-$ *DNA code* is a collection of single stranded DNA sequences of length n. In DNA hybridization assays, the general rule is that formation of WC duplexes is good, while the formation of CH duplexes is bad. A primary goal of DNA code design is be assured that a fixed temperature can be found that is well above the melting point of all CH and well below the melting point of all WC duplexes that can form from strands in the code. (It is also desirable for all WC duplexes to have melting points in a narrow range.) Thus the formation of any WC duplex must be significantly more energetically favorable then all possible CH duplexes. A DNA code with this property is said to have *high binding specificity*. High binding specificity is akin to a high signal-to-noise ratio.

A natural simplification for formulating binding specificity is to base it upon the maximum number of WC (inter-strand, non-covalent hydrogen) base pair bonds between complementary letter pairs which may be formed between two oppositely directed strands. Then for $x,y \in C$, an upper bound on this maximum number of base pair bonds that can form in the $x : \overleftarrow{y}$ duplex is the maximum length of a common subsequence to x and $\overline{y}$. In short, two single stranded DNA sequences x and y of length n can form d base pairs bonds in a duplex only if $\phi^1(x,\overline{y}) \leq n-d$. This doesn't mean that x and $\overleftarrow{y}$ will form d base pair bonds in a hybridization assay, it just says they could never form more than d base pair bonds.

If the binding specificity were solely dependent on the number of base pair bonds, then n-DNA codes constructed by using $\Phi^1(x,y)$ as the distance function could be used in hybridization assays with assured high binding specificity. This is because if $n-d$ is

large enough, then one could find a temperature that exceeds the d base pair bonding threshold of all $x:\overleftarrow{y}$ CH duplexes, but is below the melting point of each $x:\overleftarrow{\overline{x}}$ WC duplex in which n base pair bonds form.

However, while the melting point of DNA duplexes depends, in part, on the number of base pair bonds, the state of the art model of DNA duplex thermodynamics is the Nearest Neighbor Model (NN). In the NN model, thermodynamic (e.g., free energy) values are assigned to *loops* rather than base pairs. We now briefly discuss some key aspects of the NN model.

Consider two oppositely directed DNA strands $x = 5'x_1,x_2,\ldots,x_i,\ldots,x_n3'$ and $\overleftarrow{y} = 3'\overline{y_1},\overline{y_2},\ldots,\overline{y_j},\ldots,\overline{y_n}5'$ where $\overline{y_j}$ denotes the complement to base y_j. A *secondary structure* of the DNA duplex $x:\overleftarrow{y}$ is a sequence of pairs of *complementary* bases $((x_{i_r},\overline{y_{j_r}}))$ where (x_{i_r}) and $(\overline{y_{j_r}})$ are subsequences of x and $\overleftarrow{y}$ respectively. Clearly the duplex $x:\overleftarrow{y}$ can have many secondary structures. An important issue is to understand *which* secondary structure in the most energetically favorable.

The collection of complementary pairs in a given secondary structure of a duplex partitions the duplex in to pairs of substrings (or subduplexes)

that have the $(x_{i_r}),(\overline{y_{j_r}})$ and $(x_{i_{r+1}}),(\overline{y_{j_{r+1}}})$ as endpoints. For example, in the $x:\overleftarrow{y}$ duplex presented as:

$$5'x_1,x_2,\ldots,x_{i_1}3'\ *,\ldots,*5'x_{i_r},\ldots,x_{i_{r+1}}3'\ *,\ldots,*5'x_{i_k},\ldots,x_n3'$$
$$3'\overline{y_1},\overline{y_2},\ldots,\overline{y_{j_1}}5'\ *,\ldots,*3'\overline{y_{j_r}},\ldots,\overline{y_{j_{r+1}}}5'\ *,\ldots,*3'\overline{y_{j_k}},\ldots,\overline{y_n}5'$$

each pair

$$5'x_{i_r},\ldots,x_{i_{r+1}}3'$$
$$3'\overline{y_{j_r}},\ldots,\overline{y_{j_{r+1}}}5'$$

of substrings (separated by *) is an elementary substructure called a *loop* of the given

secondary structure $((x_{i_r}, \overline{y_{j_r}}))$ of the given duplex $x : \overleftarrow{\overline{y}}$. If each of the strings in a loop are of length 2, e.g.,

$$\begin{matrix} 5' x_{i_r}, x_{i_r+1} 3' \\ 3' \overline{y_{j_r}}, \overline{y_{j_r+1}} 5' \end{matrix},$$

then that loop is called a *stacked pair*.

Example 4. *We use mix of lower case and upper case letters to help identify the secondary structure. Consider the duplex*

$$\begin{matrix} 5' ggCaTaTcatACt 3' \\ 3' TccAAttGgtaGa 5' \end{matrix}$$

where the secondary structure is $(g,c),(g,c),(a,t),(a,t),(c,g),(a,t),(t,a),(t,a)$. *Loops are* $E_1, \ldots, E_8$ *and are listed left to right:*

$$\begin{matrix} g * gg * gGc * aTa * aTc * ca * at * tACt \\ Tc * cc * cAAt * tt * tGg * gt * ta * aGa \end{matrix}$$

$$= E_1 \underset{*}{*} E_2 \underset{*}{*} E_3 \underset{*}{*} E_4 \underset{*}{*} E_5 \underset{*}{*} E_6 \underset{*}{*} E_7 \underset{*}{*} E_8 .$$

The free energy, ΔG, *of the duplex predicted by the NN model is approximately* $\sum_{i=1}^{8} \Delta G_i$ *where* ΔG_i *is the free energy assigned to loop* E_i. *However, in many cases, the most stabilizing features of the structure come from the stacked pairs i.e.,* E_2, E_6, *and* E_7, *and the free energies of stacked pairs are the most accurately measured. See []. The free energies for most non-stacked loops are approximated from the free energy for stacked pairs with the same terminal pairs. For example, the free energy of* $E_3 = \begin{matrix} 5' gCa 3' \\ 3' cAAt 5' \end{matrix}$ *would be approximated by adding a "penalty" to the free energy for the measured free energy for the stacked pair* $\begin{matrix} 5' ga 3' \\ 3' ct 5' \end{matrix}$ *(that does not appear in the above secondary structure.) In most cases, the penalty takes on a positive value while all of the free energies for stacked pairs are negative. It is therefore reasonable to assume that if one only considers the free energies for the stacked pairs, then their sum would be*

a lower bound for the NN free energy for the given duplex with the given secondary structure.

Consider two identically directed DNA strands $x = 5'x_1,x_2,\ldots,x_i,\ldots,x_n3'$ and $y = 5'y_1,y_2,\ldots,y_j,\ldots,y_n3'$. For computational purposes, we define the idea of a *virtual secondary structure* between these two identically directed strands even thought no such structure would naturally form. A virtual secondary structure of the *virtual DNA duplex* $x:y$ is a sequence of pairs of identical bases $((x_{i_r},y_{j_r}))$ where (x_{i_r}) and (y_{j_r}) are subsequences of x and y respectively. In other words, a virtual secondary structure of the virtual duplex $x:y$ *is a common subsequence* $x_\sigma = y_\tau$ *of* x *and* y. Then the virtual duplex $x:y$ has the virtual secondary structure $((x_{i_r},y_{j_r}))$ if and only if the actual duplex $x:\overleftarrow{y}$ (where $x = 5'x_1,x_2,\ldots,x_i,\ldots,x_n3'$ and $\overleftarrow{y} = 3'\overline{y_1},\overline{y_2},\ldots,\overline{y_j},\ldots,\overline{y_n}5'$) has the actual *secondary structure* of pairs of complementary bases $((x_{i_r},\overline{y_{j_r}}))$ where (x_{i_r}) and $(\overline{y_{j_r}})$ are subsequences of x and $\overleftarrow{y}$ respectively. A stacked pair, $\begin{matrix}5'x_{i_r},x_{i_r+1}3'\\ 3'\overline{y_{j_r}},\overline{y_{j_r+1}}5'\end{matrix}$, exists in the actual secondary structure $((x_{i_r},\overline{y_{j_r}}))$ if and only if the corresponding *virtual stacked pair,* $\begin{matrix}5'x_{i_r},x_{i_r+1}3'\\ 5'y_{j_r},y_{j_r+1}3'\end{matrix}$, exists in the virtual secondary structure of the virtual duplex $x:y$. Thus, there exists a virtual stacked pair in a virtual secondary structure $x_\sigma = y_\tau$ if and only if $(x_i,x_{i+1}) = (y_{f(i)},y_{f(i)+1})$ is a common 2-string of the common subsequence $x_\sigma = y_\tau$ where $f:\sigma \to \tau$ is unique.

Identifying virtual stacked pairs with their natural representation, the *virtual free energy* (F) values can be associated to the negative of their corresponding values (ΔG) for actual stack pairs. The actual values are given in KCAL/mole measured at 37C

and with specified ionic concentrations. Table 1 gives the values with their corresponding virtual stacked pairs. Since the virtual stacked pair is a pair of identical 2-strings $(x_i,x_{i+1}) = (y_{f(i)},y_{f(i)+1})$, we can represent this virtual stacked pair by (x_i,x_{i+1}) and denote its virtual free energy by $F(x_i,x_{i+1})$. The $(i,j)^{th}$ entry of Figure 5 is the value of $F(i,j)$, e.g., $F(C,T) = 1.28$. (So $F(C,T)$ denotes the free energy associated with the $\begin{matrix}5'CT3'\\3'GA5'\end{matrix}$ naturally occurring stacked pair. We take F as our weight function on $[4^2]$.

F	A	C	G	T
A	1.00	1.44	1.28	0.88
C	1.45	1.84	2.17	1.28
G	1.30	2.24	1.84	1.44
T	0.58	1.30	1.45	1.00

Figure 5. Thermodynamic weight of virtual stacked pairs.

Let $x:\overleftarrow{y}$ be an actual duplex and let $\Delta G(x:\overleftarrow{y})$ be the NN computation of the free energy of the $x:\overleftarrow{y}$ duplex. The main point of all of this is that it is quite reasonable to assume that in most cases: $-\psi_F^2(x,y) \le \Delta G(x:\overleftarrow{y})$. From a DNA duplex point of view, with F being the thermodynamic weight of the virtual stacked pairs of nucleotides, we have: $\|x^{(2)}\|_F = -\Delta G(x:\overleftarrow{x})$. Thus if C is a F weighted stacked paired (A,C,G,T) quaternary code of distance d, then: $x \ne y \in C \Rightarrow \Psi_F^2(x,y) \ge d$. This implies that the thermodynamic stability, $-\Delta G(x:\overleftarrow{x})$ and $-\Delta G(y:\overleftarrow{y})$, of each (all) of the WC duplexes $x:\overleftarrow{x}$ and $y:\overleftarrow{y}$, respectively would each be at least "d greater than" the thermodynamic stability, $-\Delta G(x:\overleftarrow{y})$, of the (any) $x:\overleftarrow{y}$ CH duplex where $x \ne y \in C$. Thus, n-DNA codes closed under complementation $(x \in C \Leftrightarrow \overline{x} \in C)$ constructed by using $\Psi_F^2(x,y)$ as the distance function could be

used in hybridization assays with high binding specificity.

Example 5. *Given DNA sequences* $x = GTTATAGGCCGAG$ *and* $y = CGTC\ GTGTATATT$ *of length 13, consider the virtual secondary structure* $x_\sigma = y_\tau$ *with* $\sigma = [1,6]$ *and* $\tau = [5,6],[8,11]$. *We have that* $\sigma_\tau^2 = [1,1],[3,5]$ *and* $\tau_\sigma^2 = [5,5],[8,10]$. *We use lower case letters to exhibit the common subsequences that represent the virtual secondary structures represented by* $x_\sigma = y_\tau$ *:*

$$gttataGGCCGAG$$
$$CGTCgtGtataTT\ .$$

Identify $0 \equiv A$, $1 \equiv C$, $2 \equiv G$ *and* $3 \equiv T$ *and convert the DNA sequences accordingly. Then* $x^{(2)} = \mathbf{11},15,\mathbf{12},\mathbf{3},\mathbf{12},2,10,9,5,6,8,2$ *and* $y^{(2)} = 6,11,13,6,\mathbf{11},14,11,\mathbf{12},\mathbf{3},\mathbf{12},3,15$ *where the (bold faced) block isomorphic subsequence,* $x^{(2)}_{\sigma_\tau^2} \cong y^{(2)}_{\tau_\sigma^2}$, *represents the four virtual stacked pairs* gt,ta,at,ta *in the displayed virtual secondary structure* $x_\sigma = gttata = y_\tau$. *Using the* F *in Figure 5, we have that* $\left\| x^{(2)}_{\sigma_\tau^2} \right\|_F = 3.48$. *However,* $\psi^2_F(x,y) = 3.61$. *This is because the virtual secondary structure* $x_\alpha = y_\beta$ *with* $\alpha = [1,2],[10,11],[13,13]$ *and* $\beta = [2,5],[7,7]$ *depicted using lower case letters as:*

$$gtTATAGGCcgAg$$
$$CgtcgTgTATATT$$

has $\alpha_\beta^2 = [1,1],\ [10,10]$ *and* $\beta_\alpha^2 = [2,2],\ [4,4]$. *Then* $x^{(2)}_{\alpha_\beta^2} = 11,6 = y^{(2)}_{\beta_\alpha^2}$ *represents the virtual stacked pairs gt and cg in the virtual secondary structure* $x_\alpha = gtcgg = y_\beta$. *Finally, we have that* $\psi^2_\Omega(x,y) = \left\| x^{(2)}_{\alpha_\beta^2} \right\|_\Omega = 3.61.$

Example 6. *Given DNA sequences* $x = AATCCAACATTATTGC$ *and* $y = GTCACATCATCAAGCC$ *and using the* F *in Figure 5, we have* $\|x^{(2)}\|_F = 18.39,$

$\|y^{(2)}\|_F = 20.7$ *and* $\psi_F^2(x,y) = 8.19.$ *Thus* $\Psi_F^2(x,y) = 10.20.$ *We also have that* $x^{(2)} = 0,3,13,5,4,0,1,4,3,15,12,3,15,14,9$;

$y^{(2)} = 11,13,4,1,4,3,13,4,3,\ 13,4,0,2,9,5.$

There are at most six stacked pairs in any virtual secondary structure between x *and* y, *i.e.,* $\psi^2(x,y) = \phi^2(x^{(2)},y^{(2)}) = 6.$ *A virtual secondary structure that has six stacked pairs is* $x_\sigma = x_{[3,4],[7,10],[12,13],[15,16]} = y_{[2,6],[8,9],[14,15]} = y_\tau.$ *These six stacked pairs are represented by the common block isomorphic subsequence* $x^{(2)}_{\sigma^2_\tau} = x^{(2)}_{[3,3],[7,9],[12,12],[15,15]} \cong y^{(2)}_{[2,2],[4,6],[8,8],[14,14]} = y^{(2)}_{\tau^2_\sigma}.$ *In this case,* $\psi_F^2(x,y) = \phi_F^2(x^{(2)},y^{(2)}) = \left\| x^{(2)}_{[3,3],[7,9],[12,12],[15,15]} \right\|_F = 8.19.$ *We also have that*

$x^{(3)} = 2,9,36,20,16,1,4,18,10,39,33,10,42,45$

$y^{(3)} = 57,35,17,4,18,9,35,18,9,\ 35,16,3,13,53.$

Since $\psi^3(x,y) = \phi^3(x^{(3)},y^{(3)}) = 2,$ *we have that most number of 3-stems in any secondary virtual secondary structure between* x *and* y *is* $2.$ *Note that* $x^{(3)}_{[7,8]} \underset{3}{\cong} y^{(3)}_{[4,5]}.$ *Note that the virtual secondary structure* $x_\sigma = x_{[3,4],[7,10],[12,13],[15,16]} = y_{[2,6],[8,9],[14,15]} = y_\tau$ *has exactly two 3-stems, namely* $x_{[7,9]} = y_{[4,6]} = ACA$ *and* $x_{[8,10]} = y_{[5,7]} = CAT.$

5.2 Biomolecular Computing Architecture

A DNA bit string of length N is a DNA molecule (single long strand) that consists of N distinct nonoverlapping substrands. Suppose we have a DNA code C of size 4N partitioned into coding strands (2N) and probe stands (2N). For example, consider the DNA code below. It has twenty codewords strands each 12 bases long. Ten of these are labelled T_i or F_i and ten are labelled Probe(T_i) or Probe(F_i). Probe(X) is the WC

complement of X. This allows use to code and read 32 DNA bit strings. The DNA library has 32 longer strands of 60 bases of the form $X_1 X_2 X_3 X_4 X_5$ where $X_i = T_i$ or F_i as given below. See Figure 6.

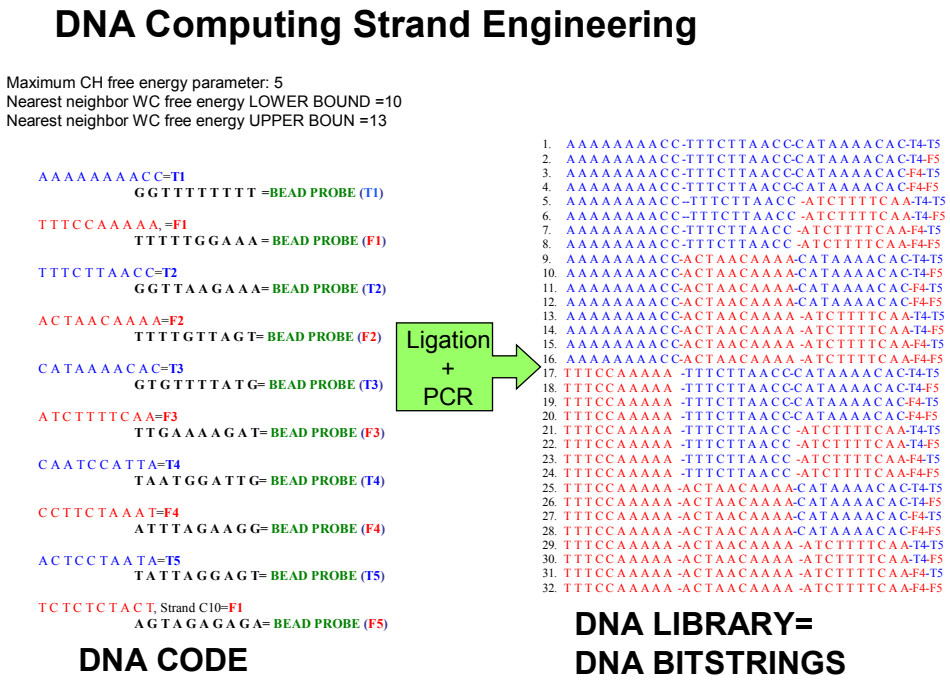

Figure 6. DNA Strand Engineering

As indicated above, we identify DNA bit strings and binary sequences. For $I \subseteq [N]$ and $(e_i)_{i \in I}$ a binary sequence, let K be the following a subset of binary N-sequences defined as $K = \{(b_i) : b_i = e_i \text{ for some } i \in I\}$. K is the set of all binary sequences that satisfy the disjunctive clause K' over N Boolean terms, each of which is a variable x_i (if $e_i = 1$) or its negation $\sim x_i$ (if $e_i = 0$.) The main "computing" idea is an iteration of the following: Given a subset T of DNA bit strings and a set K defined above, the subset $T \cap K$ can be extracted from the set T by hybridization.

We now discuss a problem that is of particular interest to us.

Problem 1. Let $P_1, P_2, ..., P_m$ be fixed subsets of [N].

a. Find all $S \subset [N]$ with $S \not\subset P_i$ for all i with $1 \le i \le m$.

b. Find all $T \subset [N]$ with $P_i \not\subset T$ for all i with $1 \le i \le m$.

Both of these problems are related and are simplified forms of the general SAT problem. They can be solved by the method described above. (These are simplifications because no negations appear in the clauses.)

There is one important difference. In the SAT problem, only one solution needs to be found. Here all solutions are required.

Let (b_i) be a binary n-sequence. As above, let $K_i = \{(b_j) : b_j = 1 \text{ for some } j \notin P_i\}$. Clearly all $S \not\subset P_i$ for all i with $1 \le i \le m$ is the set of all S with incidence vector in $\bigcap_{i=1}^{m} K_i$. In the DNA bit string representation, $K_i = \{(b_j) : b_j = r_j \text{ for some } j \notin P_i\}$. The associated filter F_i consists of $\{l_j : j \notin P_i\}$. If a set S of DNA bit strings of length N is passed through F_i, then only the bit strings in K_i remain in the gel of F_i. Starting with all possible DNA bit strings and iterating the filter process outlined above m times, we arrive at F_m. F_m contains all the DNA bit string representations of the solutions to Problem 1a. Problem 1b can transformed into Problem 1a because $P_i \not\subset T$ if and only if $[N] - T \not\subset [N] - P_i$.

The most straightforward application of the above problem is in the identification of *independent sets* in a graph. If one takes all the edges of a simple graph G as the collection $\{P_i\}$, then the set of all T is the collection of independent sets in G. See Figures 7-10.

DNA Computing for Independent Sets

Let $Q_1, Q_2, ...Q_k$ be fixed subsets of $\{1,2,...,n\}$.

a. Find all subsets $S \subseteq \{1,2,...,n\}$ with $S \not\subset Q_i$ for i with $1 \le i \le k$.
b. Find all subsets $T \subseteq \{1,2,...,n\}$ with $Q_i \not\subset T$ for i with $1 \le i \le k$

Let $\{1,2\},\{1,4\},\{2,3\},\{2,5\},\{3,4\},\{4,5\}=Q_1,...,Q_6$ be fixed subsets of $\{1,2,...,5\}$.
Finding all subsets $T \subseteq \{1,2,...,n\}$ with $Q_i \not\subset T$ for i with $1 \le i \le 6$, is finding all independent sets in G or all cliques in the complement G'.

Figure 7. Independent Sets Problem

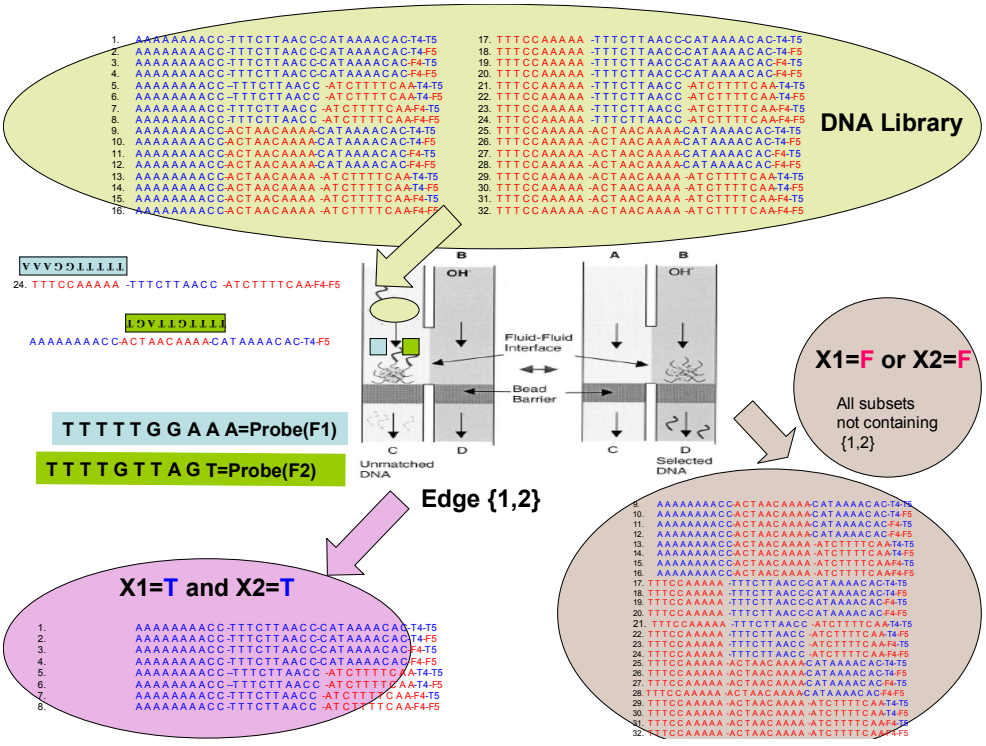

Figure 8. Biomolecular Edge Filter

The filter above give all sets of vertices that do not contain edge {1,2}. If this process is iterated (as in Figure 8), all independent sets (or cliques) will be identified.

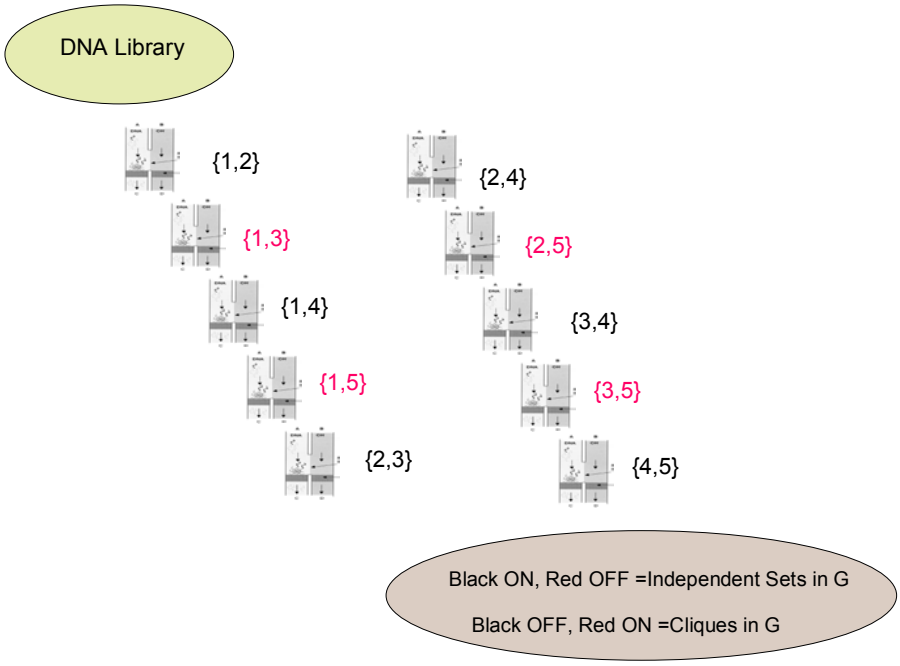

Figure 9. Filters for Edges of Graph in Figure 7.

Outflow from previous (red or black, but not both) filter is passed on to the next filter of the same color. The final outflow is the set of molecules that represent independent sets (black) or cliques (red). This system is for the graph(s) in Figure 7. A universal system is described in Figure 10.

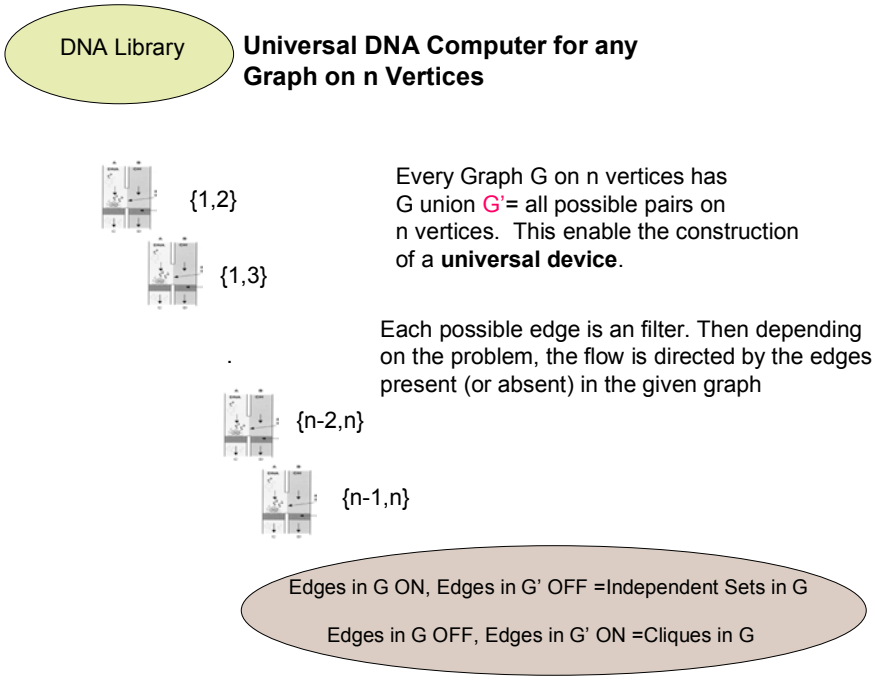

Figure 10. Universal DNA Independent Set Computer

5.3 t-Stem DNA Code Generation Software

We describe a program which we make freely available. The program(s) generates DNA codes. Some of the inputs are:

1. Length of DNA codewords: n ; 2. Stem sizes checked: $t_1, t_2, \cdots$; 3. Corresponding thresholds for each stem size: $s_1, s_2, \cdots$; 4. Maximum CH free energy parameter: ΔG_{CH} ; 5. Nearest neighbor WC free energy lower bound parameter: 6. Nearest neighbor WC free energy upper bound parameter: $\overline{\Delta G_{wc}}$.

What is generated is a DNA code C such that:

1. $x \in C \Rightarrow |x| = n$ and $\overline{x} \in C$. Thus the WC complement of each strand in the code is also in the code.
2. $x \neq y \in C \Rightarrow \psi^{t_i}(x,y) \leq s_i$. Thus the maximum number of $t_i - stems$ in each CH duplex from C is at most s_i.
3. $x \neq y \in C \Rightarrow \psi_F^2(x,y) \leq \Delta G_{CH}$. Thus each CH duplex in C has a free energy of formation above $-\Delta G_{CH}$
4. $x \in C \Rightarrow \underline{\Delta G_{wc}} \leq \|x^{(2)}\|_F \leq \overline{\Delta G_{wc}}$. Thus each WC duplex in C has a free energy of formation between $\overline{-\Delta G_{wc}}$ and $\underline{-\Delta G_{wc}}$.

Example 7. *Below is a DNA code generated by one of our programs with the inputs:* $n=16$; $t_1, t_2, t_3 = 1,2,3$; $s_1, s_2, s_3 = 10,6,2$; $\Delta G_{CH} = 8$; $\underline{\Delta G_{wc}} = 18$; $\overline{\Delta G_{wc}} = 22$. *No codeword contains GGG or CCC as a substring. The complement of any strand is either to the immediate right or left of the given strand. There are 30 codewords in the code below.*

GGCCAAAAAAAAAAAA, TTTTTTTTTTTTGGCC, GGCAAAGGTTTTCCAA,
TTGGAAAACCTTTGCC, CATTTTAAGGAACCGG, CCGGTTCCTTAAAATG,
TCCTCTTTCTTTACCA, TGGTAAAGAAAGAGGA, TAGAATCCGTCAATTT,
AAATTGACGGATTCTA, GGTTACGGTGGTGTTT, AAACACCACCGTAACC,
TTTGTCACTTGTGGAG, CTCCACAAGTGACAAA, AGTATTTCGATCTTCC,
GGAAGATCGAAATACT, CAGGCGTTGATGAACA, TGTTCATCAACGCCTG,
TAACTATGTAGCATGG, CCATGCTACATAGTTA, CAACAATAGGAGGCTT,
AAGCCTCCTATTGTTG, GGACTTAGGCAGACGT, ACGTCTGCCTAAGTCC,
GAGCGAGGTAGATTAG, CTAATCTACCTCGCTC, GATACACACGGCATAT,
ATATGCCGTGTGTATC, CGAGTGGCTCTCTCAT, ATGAGAGAGCCACTCG,

To further minimize errors in the applications, further constrains on the code were considered. Below is a DNA code generated by one of our programs with the inputs: n=12; t_1 =2; s_1 =6; ΔG_{CH} =7 ; $\underline{\Delta G_{wc}}$ = 14; $\overline{\Delta G_{wc}}$ = 17. No codeword contains GGG or CCC as a substring. In addition, in any given WC pairs of DNA codewords, only one strand contains a G. This is achieved by selecting "ACT-AGT only" This strand will be used as the probe strand. Thus the WC complement of strand X is listed as Probe(X) below. Moreover, with the addition of the coupling constraint we also ensure that sequences that are formed in the middle of the junctions of any library strand $T_i T_{i+1}$, $T_i F_{i+1}$, $F_i T_{i+1}$, $F_i F_{i+1}$ all obey the code constraints. This is to ensure that probes do not hybridize at locations where code strands are ligated into library strands. See Figure 10.

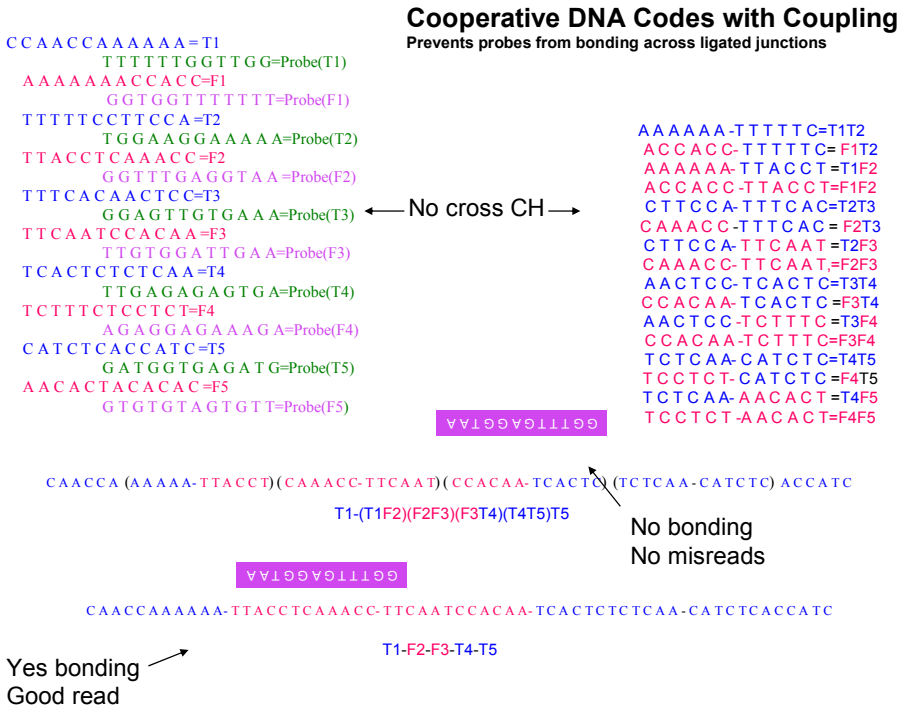

Figure 11. Cooperative DNA Code

6. References

1. M. Andronescu, A. Condon and H. Hoos, RNAsoft, submitted to NAR for the web-based software special issue, available at http://www.rnasoft.ca/

2. M. Andronescu, Algorithms for predicting the secondary structure of pairs and combinatorial sets of nucleic acid strands, Masters Thesis, University of British Columbia, (2003.)

3. E. Baum. DNA sequences useful for computation, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 44, 235-242, (1999.)

4. Braich, R., Chelyapov, N., Johnson, C., Rothemund, P.W.K., Adleman, L. Solution of a 20-Variable 3-SAT Problem on a DNA Computer. Sciencexpress, 1-15, (2002).

5. A. Brenneman and A. Condon, Strand Design for biomolecular computation, Theoretical Computer Science, 287, 39-58, (2002).

6. H. Cai, et al., Flow Cytometry-Based Minisequencing: A New Platform for High Throughput Single Nucleotide Polymorphism Scoring, Genomics, 66, 135-143, (2000.)

7. A. D'yachkov and D. Torney, On Similarity Codes, IEEE Trans. on Information Theory 46, 1558-1564, (2000.)

8. R. Deaton, et al., A PCR Based Protocol for in Vitro Selection of Noncrosshybridizing Oligonucleotides, DNA Computing, DNA 8, M. Hagiya, A. Ohuchi (eds.), LNCS 2568, Springer, Berlin 196-204 (2002.)

9. R. Deaton, et al., A Software Tool for Generating Noncrosshybridizing Libraries of DNA Oligonucleotides, DNA Computing, DNA 8, M. Hagiya, A. Ohuchi (eds.), LNCS 2568, Springer, Berlin 252-261 (2002.)

10. A. D'yachkov, et al., On a Class of Codes for Insertion-Deletion Metric, 2002 IEEE Intl. Symp. Info. Th., Lausanne, Switzerland, (2002.)

11. A. D'yachkov, et al., Exordium for DNA Codes, Journal of Combinatorial Optimization, 7, no.4, 369-380 (2003.)

12. A. D'yachkov, et al., Reverse-Complement Similarity Codes, IEEE Trans.on Information Theory to appear

13. P. Erdos, D. Torney, and P. Sziklai, A Finite Word Poset, Elec. J. of Combinatorics, 8, (2001.)

14. M. Garzon, et al., A new metric for DNA computing, in Genetic Programming 1997: Proceedings of the Second Annual Conference, pp. 479-490, AAAI, 1997. Stanford

University, July 13-16, 1997.

15. D. Gusfield, Algorithms on Strings, Trees, and Sequences, Cambridge, (1997.)

16. Hartemink, A., Gifford, D., A thermodynamic simulation of deoxyoligonucleotide hybridization for DNA computation, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 48, 25-37 (1999.)

17. H. Hollman, A relation between Levenshtein-type distances and insertion and deletion correcting capabilities of codes, IEEE Trans. on Information Theory, 39 1424-1427, (1993.)

18. V. Levenshtein, Efficient reconstruction of sequences from their subsequences or supersequences, Journal of Combinatorial Theory, Series A, 93, 310-332 (2001.)

19. V. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, Soviet Phys.—Doklady, 10 707-710, (1966).

20. V. Levenshtein, Bounds for Deletion-Insertion Correcting Codes, 2002 IEEE Intl. Symp. Info. Th., Lausanne, Switzerland, (2002).

21. A. Macula, DNA-TAT Codes, USAF Technical Report, TR-2003-57, AFRL-IF-RS http://stinet.dtic.mil/cgi-bin/fulcrum_main.pl (2003.)

22. A. Macula, et al., DNA Code Gen., available at https ://community.biospice.org

23. J. SantaLucia Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. USA, Vol. 95, pp 1460-1465 (1998.)

24. M. Waterman, Introduction to Computational Biology, Chapman-Hall, London,

(1995.)

25. A. Zuker, B. Mathews and C. Turner, Algorithms and Thermodynamics for RNA Secondary Structure Prediction: a Practical Guide